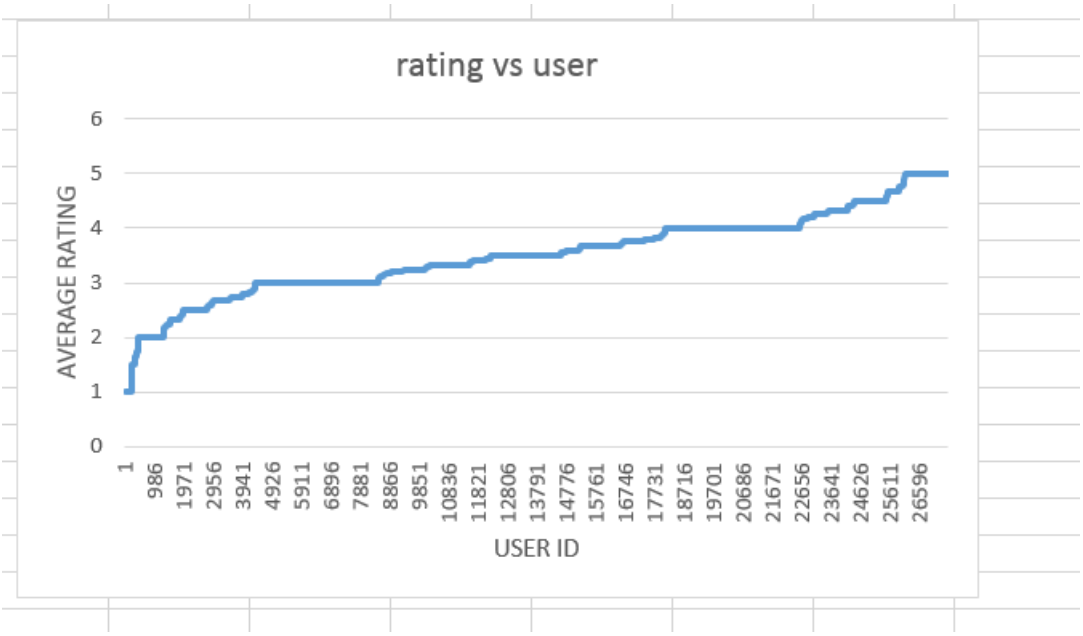
Q1， the tpo 15 movie after linking them with name is

12293	The Godfather
4432	The Italian Job
4640	Rain Man
6408	Good Morning
8596	Seven
9728	As Good as It Gets
13651	Air Force One
1744	Beverly Hills Cop
1202	National Lampoon's Vacation
13614	Office Space
2660	When Harry Met Sally
6287	Pretty Woman
10947	The Incredibles
8915	Terminator 2: Extreme Edition
6971	Ferris Bueller's Day Off

Q1 part b

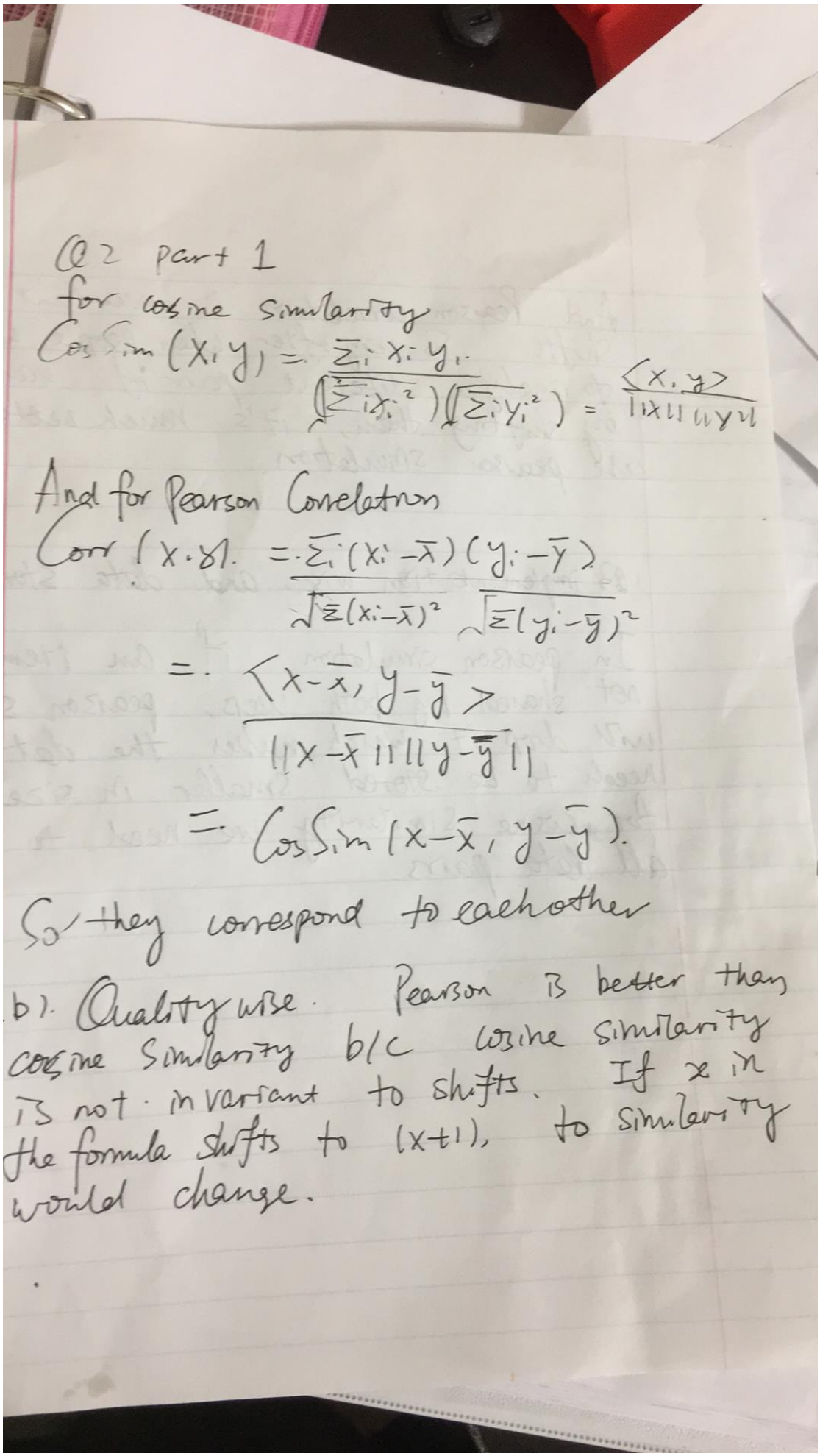
The distribution of average rating is



The fraction of overly enthustic customer is 5406/27555= 0.196189

I think I should remove the overly enthusiastic customers to make my result less biased.

Q2



Q2 part 1

for cosine similarity

$$CosSim(X,Y) = \frac{\sum_i x_i y_i}{\sqrt{(\sum_i x_i^2)}\sqrt{(\sum_i y_i^2)}} = \frac{\langle x, y\rangle}{\|x\|\|y\|}$$

And for Pearson Correlation

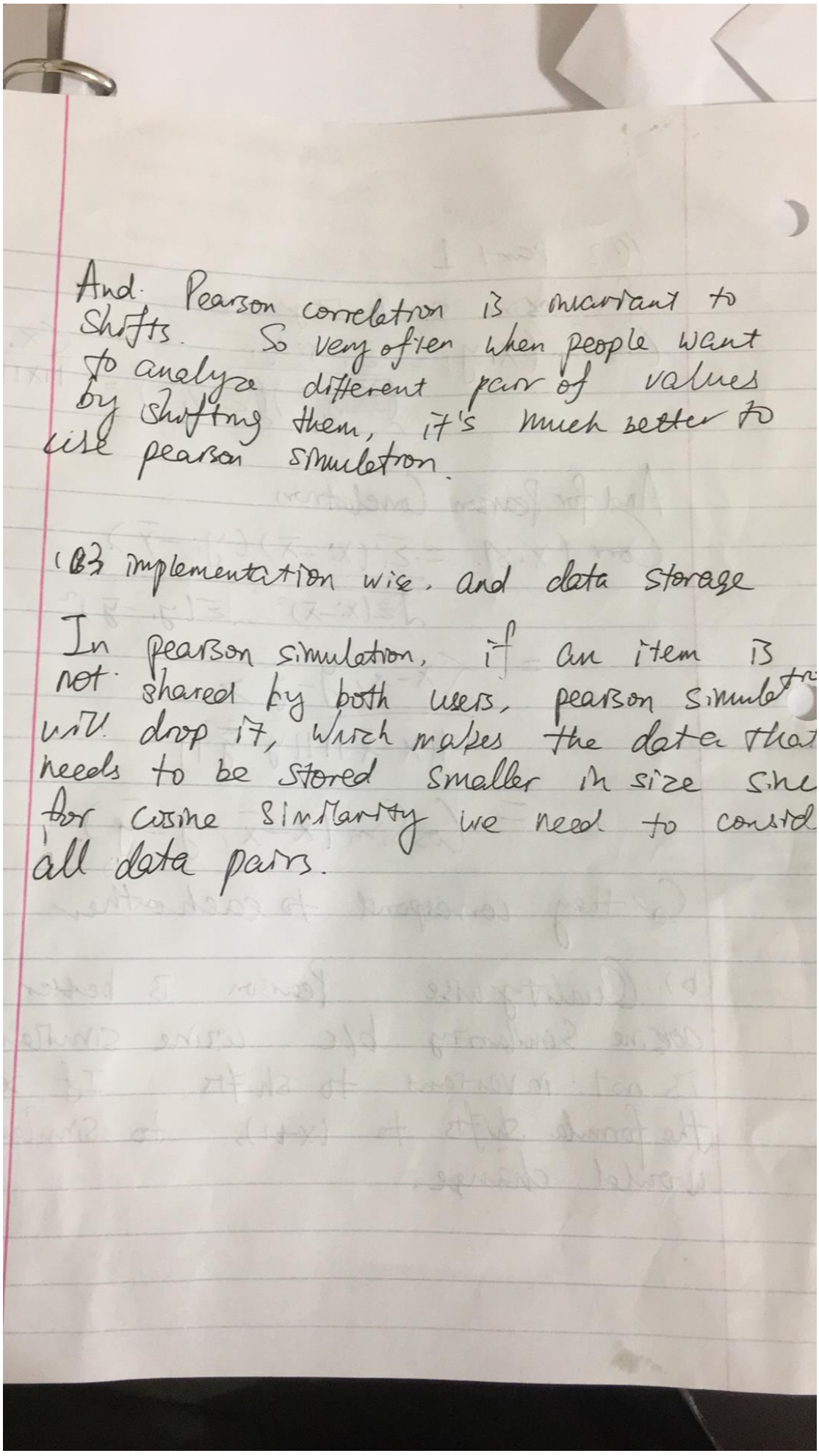
$$Corr(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$$

$$= \frac{\langle x-\bar{x}, y-\bar{y}\rangle}{\|x-\bar{x}\|\|y-\bar{y}\|}$$

$$= CosSim(x-\bar{x}, y-\bar{y}).$$

So they correspond to eachother

b). Quality wise. Pearson is better than cosine similarity b/c cosine similarity is not invariant to shifts. If x in the formula shifts to (x+1), to similarity would change.



And. Pearson correlation is invariant to shifts. So very often when people want to analyze different pair of values by shifting them, it's much better to use pearson simuletion.

(B) implementation wise, and data storage

In pearson simulation, if an item is not shared by both users, pearson simulation will drop it, which makes the data that needs to be stored smaller in size. The for cosine similarity we need to consid all data pairs.

Partc

I think Jaccard ignores the value of individual elements in the set, and it would take longer to calculate what A and B both have and not have.

So I think it would be more efficient to preprocess the data to get the actual value of it and then calculate jaccard measure, it would be more accurate.

Job1.
Mapper output
(user1,< movie1,1>)
(user1,< movie3,2>)
(user1,< movie2,3>)
(user2,< movie2,2>)
(user2,< movie3,3>)
(user2,< movie5,5>)

Reducer input
(user1,< movie1,1>)
(user1,< movie3,2>)
(user1,< movie2,3>)
(user2,< movie2,2>)
(user2,< movie3,3>)
(user2,< movie5,5>)

Reducer output

(<movie1,movie2>, <1,3>)

(<movie1,movie3>, <1,2>)

(<movie2,movie3>, <3,2>)

(<movie2,movie3>, <2,3>)

(<movie2,movie5>, <2,5>)

(<movie3,movie5>, <3,5>)

2nd step
Mapper output

(<movie1,movie2>, <1,3>)

(<movie1,movie3>, <1,2>)

(<movie2,movie3>, <3,2>)

(<movie2,movie3>, <2,3>)

(<movie2,movie5>, <2,5>)

(<movie3,movie5>, <3,5>)

2nd step reducer output

(<movie1,movie2>, 0.33)

(<movie1,movie3>,0.50)

(<movie2,movie5>, 0.01)
(<movie3,movie5>, 0.07)

(<movie2,movie3>, 0.04)